10/01/2020

MVA ASSIGNMENT 3

MEMBER INFORMATION

Ruixin Yang (RUID: 197000459)

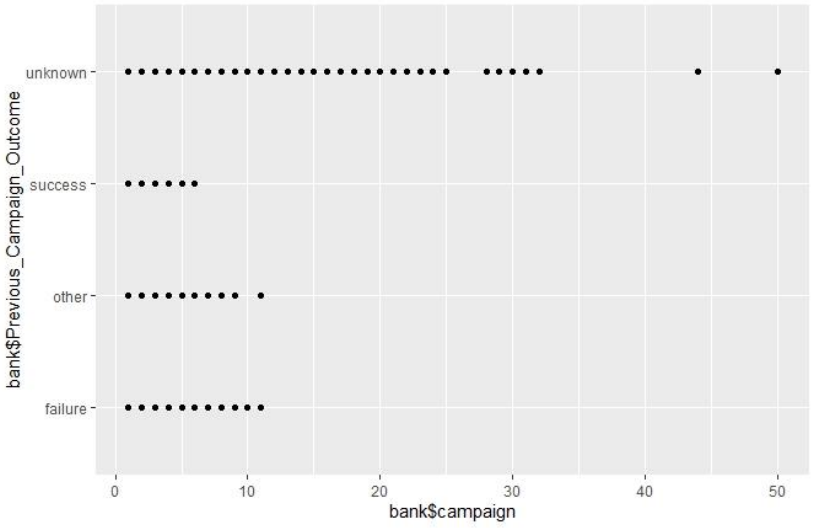
Aishwarya Senthilvel (RUID: 199001269)

DATA CLEANING & EXPLORATORY DATA ANALYSIS

The data set we have obtained is predominantly clean. However, we have performed various operations to better clean the data.

1. GGPLOT

ggplot() function was used to understand the layers of Previous_Campaign_Outcome column that holds categorical values , namely Failure, Success, Other and Unknown.
This field gives insights on the outcome of the previous marketing campaign. There are outliers found on the unknown value while the other three values are well within the observed range.
Below is the result:

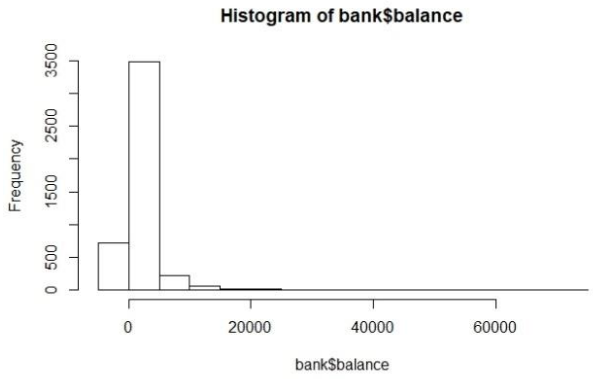


2. HISTOGRAM

The hist() function wS used to analyse frequency of bucketed values of the variable “balance” which displays the bank balance of an account holder which can help analyse the potential of that specific account to purchase a loan from the bank.

There is a high frequency of 3500 spotted at bucket 2 of account bank balance.

Below is the result:



3. DATA SUMMARY AND IDENTIFYING NULL VALUES

- Summarizing the data using summary() function enables us to view a simplified data set to observe what is normal and what looks odd.

```
> summary(bank)
      age            job              marital           education         Default_Credit
 Min.   :19.00   Length:4521        Length:4521        Length:4521        Length:4521
 1st Qu.:33.00   Class :character   Class :character   Class :character   Class :character
 Median :39.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :41.17
 3rd Qu.:49.00
 Max.   :87.00
    balance        housing             loan             contact              day
 Min.   :-3313   Length:4521        Length:4521        Length:4521        Min.   : 1.00
 1st Qu.:   69   Class :character   Class :character   Class :character   1st Qu.: 9.00
 Median :  444   Mode  :character   Mode  :character   Mode  :character   Median :16.00
 Mean   : 1423                                                           Mean   :15.92
 3rd Qu.: 1480                                                           3rd Qu.:21.00
 Max.   :71188                                                           Max.   :31.00
    month              duration        campaign          pdays            previous
 Length:4521        Min.   :   4   Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000
 Class :character   1st Qu.: 104   1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000
 Mode  :character   Median : 185   Median : 2.000   Median : -1.00   Median : 0.0000
                    Mean   : 264   Mean   : 2.794   Mean   : 39.77   Mean   : 0.5426
                    3rd Qu.: 329   3rd Qu.: 3.000   3rd Qu.: -1.00   3rd Qu.: 0.0000
                    Max.   :3025   Max.   :50.000   Max.   :871.00   Max.   :25.0000
 Previous_Campaign_Outcome Subscription_To_Deposit
 Length:4521               Length:4521
 Class :character          Class :character
 Mode  :character          Mode  :character
```

In this high-level overview we notice that all the fields are within consistent values.

Below is the result:

- We have used the function complete.cases() that identifies null values in the table. Our dataset has no null values.

  Below is the compressed result:

[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

[685] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

[704] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRU…

4. CENTRAL TENDENCIES & VARIANCE

We have calculated mean, median, and standard deviation on a key parameter called “Previous” which keeps track of the number of contacts performed before the current campaign and for the specific client.

Below is the result:

```
> mean(bank$previous)
[1] 0.5425791
> median(bank$previous)
[1] 0
```

```
> sd(bank$previous)
[1] 1.693562
```

5. SUBSET TO FIND COLUMN-WISE MISSING VALUE

We use the subset() function to identify if the “loan” column that determines if a specific customer account has a personal loan is null.

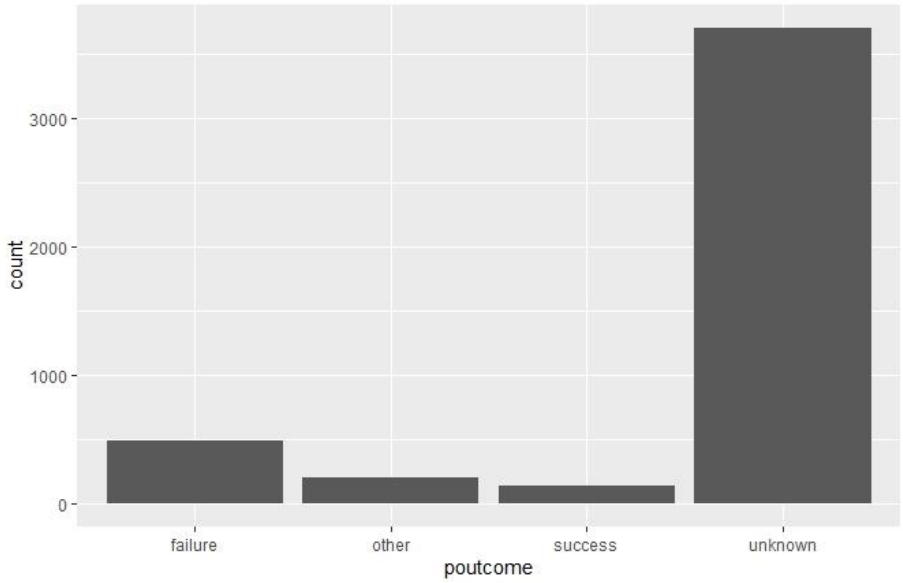
The function works, but since there is no null values in loan column it throws no output.

```
#Creates subset of bank data where loan column is missing
bank_1 <- subset(bank, is.na(bank$loan)) # this works. But there is no missing value in the column.
```

6. BARPLOT

Generating Barplot on categorical column “Poutcome” to analyze outcome of the previous marketing campaign. High frequency lies in value ‘unknown’.

Below is the result:



10/01/2020

TESTS

I summarize these tests in the last lecture. There is t test, F test, Hotelling T2 test, Levene's Test, Van Valen Test, Box M test, ANOVA test, MANOVA test, and the Mantel test. Since we want to know factors driving clients to buy financial products from banks, we will mainly focus on the differences on many variables between clients who bought the term deposit product and people who did not.

Our dataset contains the variables and their symbols presented in the data dictionary in previous assignment. In total, there are 21 variables.

1. T TEST

First, we ran a t test to see if there is any difference in clients' age between clients who bought this term deposit and clients who did not. Here is the result.

```
Two Sample t-test

data:  age[y == "yes"] and age[y == "no"]
t = 6.1721, df = 41186, p-value = 6.802e-10
alternative hypothesis: true difference in means is not equal to
 0
95 percent confidence interval:
 0.6837762 1.3201463
sample estimates:
mean of x mean of y
 40.91315  39.91119
```

You can see that the p-value is quite small here. Under the significance level of 5%, we reject the null hypothesis. So, there is a difference in age between those clients.

2. HOTELLING T^2 TEST

Then we ran a Hotelling T^2 test to see if there is any difference on means of some numeric variables between clients who bought the term deposit and clients who did not. The result is the following.

```
> print(t2testsparr)
Test stat:  2426.6
Numerator df:  8
Denominator df:  41179
P-value:  0
```

As you can see, the p-value is close to zero again. So, we choose to reject the null hypothesis. That means there is significant differences on these means between these clients.

3. F TEST

Then I conduct a F test for the variances of "duration variable" on two types of clients mentioned before. Here is the result.

```
F test to compare two variances
```

```
data:  duration[y == "yes"] and duration[y == "no"]
F = 3.7525, num df = 4639, denom df = 36547, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
 3.595103 3.919580
sample estimates:
ratio of variances
          3.752462
```

The “duration” variable represents the last contact duration in seconds. And we test for the variances of it within two groups. The p-value is again very small. With the significance level of 5%, we reject the null hypothesis. Thus, there exists a difference between those two groups of clients.

4. LEVENE’S TEST (Absolute Difference)

After that, we applied Levene’s test for absolute differences by using t-test. Here is the result.

```
Two Sample t-test

data:  absall$campaign[y_n == 1] and absall$campaign[y_n == 0]
t = -4.4704, df = 41186, p-value = 3.913e-06
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.107538
sample estimates:
mean of x mean of y
 1.273276  1.443417
```

As you can see, we want to know if there is difference in means of the “campaign” variable between these two groups of clients (1 means people bought it. 0 means people who did not). This variable represents the number of contacts performed during this campaign and for this client. It is numeric and includes last contact. The test shows that we should reject the null hypothesis since the p value is smaller than 0.05. It means that there exist differences in the variables between these two groups of clients.

5. LEVENE’S TEST (Homoskedasticity)

We then apply the levene’s test for homoskedasticity. We want to know if there is a difference between variances of “campaign” in these two groups. The result is the following.

```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     1  127.71 < 2.2e-16 ***
      41186
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As you can see, the p value here is smaller than 0.05. Thus, we reject the null hypothesis. Homoskedasticity is denied here.

6. VAN VALEN'S TEST

We next conduct a Van Valen's test for the d values between these two groups. This d value is gotten by taking square root of sums of rows in the matrix that each element is squared. Here is the result.

```
        Two Sample t-test

data:  dij[y_n == 1] and dij[y_n == 0]
t = 6.564, df = 41186, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 32.01847
sample estimates:
mean of x mean of y
 219.7881  194.1854
```

As you can see, the Van Valen's test fails to reject the null hypothesis. That means the true difference in d values is less than zero.

7. ANOVA TEST

We then apply the ANOVA test for the difference in means of "campaign" variable between these two groups. Here is the result.

```
Df Sum Sq Mean Sq F value Pr(>F)
y              1   1392  1391.6   182.2 <2e-16 ***
Residuals   41186 314635     7.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see, the p value here is smaller than $2*10^{-16}$. So, we reject the null hypothesis. We think that there is a difference in "campaign" variable between these two groups.

8. MANOVA TEST

Later, we use the MANOVA test for the differences in multiple variables between these two groups. The result is the following.

```
Df Pillai approx F num Df den Df    Pr(>F)
y_n          1 0.3231   2183.9      9  41178 < 2.2e-16 ***
Residuals 41186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see here, the p-value is still smaller than 0.05. That means there exist differences between these two groups.

9. BOX M TEST

We next use the Box M test for the homogeneity between variance-covariance matrices of these two groups. That is for variables including duration, campaign, pdays, previous, cons.price.idx, euribor3m, and nr.employed. Here is the result.

```
Box's M-test for Homogeneity of Covariance Matrices

data:  vars
Chi-Sq (approx.) = 43358, df = 45, p-value < 2.2e-16
```

As you can see, the p value is smaller than 0.05. Thus, we reject the null hypothesis. There is no homogeneity between matrices.